Steps on the Way to Equilibrium

Wayne C. Myrvold
Department of Philosophy
The University of Western Ontario
wmyrvold@uwo.ca

Forthcoming in Daniel Bedingham, Owen Maroney, and Christopher Timpson, eds., *Quantum Foundations of Statistical Mechanics* (Oxford University Press).

Abstract

A shift in focus, of the sort recently advocated by David Wallace, towards consideration of work in nonequilibrium statistical mechanics has the potential for far-reaching consequences in the way we think about the foundations of statistical mechanics. In particular, consideration of the approach to equilibrium helps to pick out appropriate equilibrium measures, measures that are picked out by the dynamics as "natural" measures for systems in equilibrium. Consideration of the rationale for using such measures reveals that the scope of their legitimate employment is much more limited than an appeal to a Principle of Indifference would suggest. These points are illustrated by use of a toy model that I call the *parabola gadget*.

Contents

1 Introduction

In recent work, David Wallace (2015, forthcoming) has directed the attention of philosophers working on the foundations of statistical mechanics to the rich array of techniques and results in nonequilibrium statistical mechanics. In my opinion, this move has the potential to be far-reaching, as it can shed new light on a deep question in the foundations of statistical mechanics, namely, the question of how we are to think of the probability measures invoked in statistical mechanics.

There has been a tendency in the literature to focus attention on the fact that systems out of equilibrium tend to equilibrate, and to offer explanations of this fact that invoke the fact for certain systems, the equilibrium macrostate is vastly larger in microcanonical measure than any nonequilibrium macrostate. Even if an explanation of equilibration could be extracted from this observation (and considerable work is required to do so), the task would remain of explaining, not merely the end-point of the process of equilibration, but the steps along the way.

Moreover, there are systems that refuse to relax to a state in which measurable parameters remain constant. Consider, for example, a Brownian particle suspended in a fluid. Take the velocity of the particle as an observable quantity that we use, along with other parameters, to characterize the state of the system. Then the equilibrium condition is not an equilibrium macrostate in which these parameters have constant values, but rather, a condition in which the velocity of the particle fluctuates according to a well-defined probability distribution. Nonetheless, we have relaxation towards that equilibrium condition. If, for example, the Brownian particle is introduced into the fluid with some speed that is greater than its mean equilibrium speed, it will tend to slow down, due to friction with the fluid, and approach a condition in which the velocity is well represented by the equilibrium distribution. This process is governed by the *Langevin equation*.[1]

As Wallace has emphasized, explanation of the success of statistical mechanics requires explanation of the success of the Langevin equation, Fokker-Planck equation, and the like. These are stochastic equations, yielding probabilities for future values of certain variables in terms of either their present values or the history of these values. Though, in principle, the future value of *any* variable depends on the

[1]See standard textbooks of nonequilibrium statistical mechanics, *e.g.* Zwanzig (2001, §1.2), Mazenko (2006, §1.2).

full state of the system, it turns out that, for a wide variety of systems, we can focus on a few variables and get autonomous equations in terms of those variables alone. Those seeking to explain the success of statistical mechanics should be in a position to explain, not only the qualitative fact of the approach to equilibrium, but also the success of methods such as this used to track the evolution of the system towards equilibrium.

In what follows, I will make some suggestions as to how this might work, by reference to a toy example for which it provably *does* work. First, though, some remarks about the notion of probability in statistical mechanics.

2 The grand temptation

As mentioned, what is to be explained are probabilistic equations of motion for macroscopic parameters. Therefore, a few words are in order about how to think about probabilities in physics.

There's a persistent temptation to think that the probability of an event can be *defined* as the ratio of the number of ways that the event can occur to the number of ways that the world could be. In his *Philosophical Essay on Probabilities* (1814), the great Laplace comes close to succumbing to this temptation. Therein we find, as the First Principle of the calculus of probability,

> *First Principle.*—The first of these principles is the definition itself of probability, which, as has been seen, is the ratio of the number of favorable cases to that of all the cases possible.

He immediately steps back from the precipice of folly, though.

> *Second Principle.*—But that supposes the various cases equally possible. If they are not so, we will determine first their respective possibilities, whose exact appreciation is one of the most delicate points of the theory of chance.

In this little dialogue we find, in miniature, a foreshadowing of much of the subsequent discussions of the Principle of Indifference: temptation to regard it, or something like it, as the foundation of probability theory, followed by critics who warn, correctly, that, unless supplemented by some judgment about *which* partition of events we are to declare equiprobable, it serves as *no guide at all*.

When the space of possibilities is a continuum, the temptation is to define the probability of an event as the ratio of the measure of the set of ways the event could occur to the measure of all the ways that the world could be. The same problem arises; choosing a measure is equivalent to specifying which sets of events are equiprobable. One might interpret the principle of indifference as enjoining us to choose a measure as one that has a flat density function, but, as is well known, a density function that is flat when written in terms of one set of variables will not remain flat under change of variables. If equal intervals of a variable x have equal probability, this does not hold (for example) for x^2, and *vice versa*.

All of this is well-known, and has been for some time. Jeremy Butterfield has, appropriately, referred to the Principle of Indifference as "that notorious dead horse of the philosophy of probability" (Butterfield, 1996, 212). Most people, these days, would readily acknowledge that it is an illusion that probabilities can be defined, without further ado, as ratios of possibility-counts.[2] It is an illusion reminiscent of the old illusion of Rationalism, that is, the idea that Pure Thought, without empirical input, can yield substantive knowledge about the world.

Yet its influence has not entirely been shaken off. Its influence lingers on in discussions of the foundations of statistical mechanics, in the idea that a Principle of Indifference uniquely singles out a privileged class of probability measures. These privileged measures are uniform in phase space variables, or else, as uniform as possible, subject to certain macroscopically definable constraints. We should be asking (as did Gibbs), what is special about a measure that is uniform in these variables, rather than some others?

Its influence lingers in approaches to statistical mechanics, such as the neo-Boltzmannian approach,[3] that acknowledge the multiplicity of measures, and hence that a choice of measure must be made in order to apply Indifference, but nonetheless insist that there is some measure that is privileged as a typicality measure, a measure introduced by brute fiat, with no intrinsic connection to the physics of the system in question. And it lingers in the philosophy of cosmology, in the fine-tuning problem, and in any argument that concludes that we should regard it as surprising that the universe started out in a state so far

[2]For a particularly emphatic rejection of the Principle of Indifference, see Albert 2000, §3.2.

[3]See Goldstein (2001) and Price (2002) for statements of views of this sort.

from thermodynamical equilibrium.

If we allow the notorious dead horse to rest in peace, what could take its place? Here, again, we find, in the history of discussions of probability, clues as to the right track to take. Bernoulli, in 1713, spoke of cases that can happen with *equal facility*.[4] Which events those were could, according to Bernoulli, in some cases be judged by symmetry considerations, but, ultimately, were to be judged *a posteriori*.

There is, I claim, a way to make sense of the idea that certain events occur equally easily, others, more or less easily. This is not based on static considerations about the structure of the space of possibilities, but on considerations of *dynamics*. These considerations have to do with sensitivity to initial conditions, sensitivity that, for the right sort of dynamics, tends to wash out differences between probability distributions over initial conditions, in that very different probability distributions over initial microstates yield virtually the same probabilities for future values of certain macroscopic variables. There will (again, for the right sort of dynamics) be probability distributions over such variables that are stable under dynamical evolution and have the status of "attractor" distributions, in that other distributions tend to approach them. For a system of this sort that has been freely evolving long enough for this convergence to take effect, we can use the attractor distribution to judge which eventualities occur with equal facility. It is considerations of this sort, I claim, rather than an appeal to a Principle of Indifference or any other considerations divorced from dynamics, that underwrite the use of standard probability distributions in statistical mechanics, and, indeed, in many situations in which we have well-defined probabilities. There is no need for a brute choice of a typicality measure with no intrinsic connection to the physics.

Considerations of this sort will yield measures that are appropriate, and in some sense *natural*, for systems that have been evolving freely long enough for the requisite washing-out of disagreements among input distributions to have taken place. This means that, in the statistical-mechanical context, they are appropriate for conditions of thermodynamic equilibrium. There is no rationale, *none whatsoever*, for regarding them as privileged probability distributions for systems that are far from equilibrium.

[4]See quotation in §8, below. For the history of locutions of this sort, which were common during the first century of the development of probability theory, see Hacking (1971).

However, for systems out of equilibrium, a case can be made for the employment of measures that are as much like the equilibrium measure as can be, subject to macroscopic constraints, provided that local equilibration has taken place (see section 10, below). In this way we will be able to obtain autonomous probabilistic equations of motion for certain quantities, equations that yield probabilities for future values of those quantities in terms of their present and past values. This, if one reflects on it, is a remarkable thing. If we characterize the state of a system by assigning values to a set of variables, the future value of each of these variables may depend on present values of *all of them*, and hence, probabilities for any future values of any variable requires, in the general case, a full probability distribution over the entire state space. Nevertheless, in nonequilibrium thermodynamics one often obtains autonomous equations of motion for certain macroscopic variables. This may seem *prima facie* mysterious, perhaps even impossible. To get a flavour of how it can happen, we will show, explicitly, how it can occur, in the case of a simple toy example I call the *parabola gadget*. The gadget will be introduced in the next section, and much of this chapter will be taken up with illustrating how the claims I wish to make about real systems are realized in this toy model.

Since the considerations we are invoking require at least partial equilibration, none of this will be of any avail in assigning a probability distribution over the initial state of the universe (if there is one). Is there anything that can do this job, once we have buried the dead horse? I think not, and that, if we are honest, we should admit that we really have no idea what to expect the early universe to be like; all we can do is collect evidence about what it was like and base our credences on that.

3 The parabola gadget

Consider the device that I call the parabola gadget, depicted in Figure 1. It consists of a board, one meter square, on which is inscribed a diagonal. Also inscribed in the square is a parabola, which touches the two bottom corners of the square, and whose peak touches the top of the square in the middle. There is a ball that starts out on the diagonal, and moves according to the following rule. From the diagonal, it heads vertically (upwards or downwards, as need be) towards the

parabola, until it reaches it. From the parabola, it heads horizontally (left or right) towards the diagonal, until it reaches it. The process then reiterates. In Figure 2a is shown one iteration of this process, and, in Figure 2b, four iterations. Suppose, now, you know that a

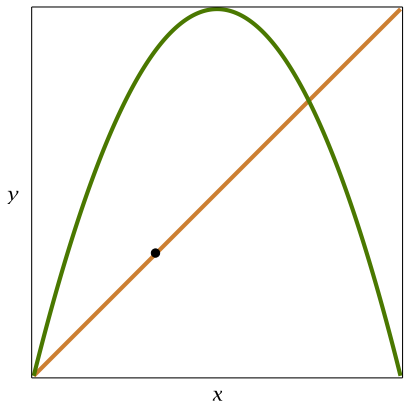
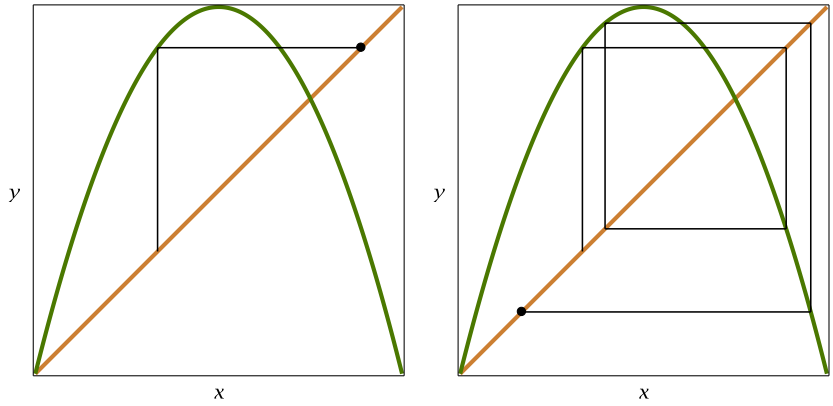


Figure 1: The parabola gadget



(a) One iteration (b) Four iterations

Figure 2: Evolution of the parabola gadget

parabola gadget has been running for some time, at least ten iterations, and that you are asked which of the two alternatives you regard as more likely (See Figure 3):

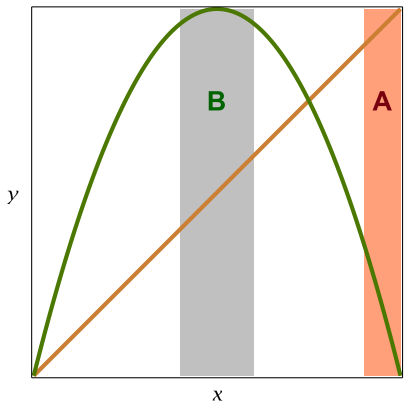


Figure 3: Alice and Bob's options

A. The ball is within 10 cm. of the right-hand side.

B. The ball is within 10 cm., on either side, of the center.

I invite you, before proceeding, to give some consideration as to which you regard as more likely. Given a choice between a reward (something you like) if A is true, and the same reward if B is true, which choice would you make?

I imagine two agents, Alice and Bob, who make different choices. Bob reasons on the basis of a Principle of Indifference, as follows.

> I know nothing about the initial conditions, and, even if I did know something, ten iterations of the gadget would render that information useless, since small differences in initial conditions can lead to large differences in outcomes. In option B the payoff conditions span a range of positions twice as large as in option A, so option B is clearly preferable.

Alice, on the other hand, regards Bob's reasoning as seriously problematic, bordering on incoherence. Here's her thinking.

> Though I am no fan of the Principle of Indifference, suppose I were to grant Bob the supposition that, at some iteration, say, the nth, I should regard intervals of equal length as equally likely. Then I can't say the same about the next iteration. It's clear from inspection of Figure 3 that all points that, at stage n, are in B, find themselves in A in

> the very next iteration. So, the ball's being in A at stage $n+1$ must be at least as likely as its being in B at stage n.
>
> Because of the shallowness of the slope of the parabola near its peak, points in some interval around the center get sent, in a single iteration of the machine, into a smaller interval near the right-hand side, which, on the next step, gets sent into a small interval near the left-hand side. There's a tendency for the ball to be more towards the edges than in the middle. On the basis of these considerations, A strikes me as more probable.

At the root of Alice's deliberations is the fact that, because of the dynamics of the machine, a probability distribution over the position of the ball at some time n uniquely determines a probability distribution over the position of the ball at time $n+1$, as follows: the probability that, at time $n+1$, the ball is in a set A is equal to the probability that, at time n, it was at some point that gets mapped into a point in A by one iteration of the gadget's evolution. In this way, given a law of evolution of the state of some physical system, we can speak of the evolution of probability distributions over its state space.

Bob's favoured distribution is unstable; applying it at some time n and also at time $n+1$ is inconsistent with what Bob knows about the dynamics of the gadget. Suppose that he applies the uniform distribution to initial conditions. On this distribution, it is more likely that the ball will be in A ten iterations down the line than that it will be in B: on a uniform measure over initial conditions, the set of states that put the ball into A after 10 iterations is larger than the set of states that put the ball into B after 10 iterations, by a factor of about 8 to 5. Both of these sets consist of a large number of small pieces, distributed over the length of the diagonal. For this reason, not only will a uniform distribution over initial conditions yield measures for these two sets that are roughly in the ratio 8/5, but the same holds for any probability distribution over initial conditions that is not "too wiggly," in a sense that can be made precise (see appendix).

It turns out that, though Bob's favoured distribution is unstable under evolution, there is another probability distribution that *is* stable.[5] Its density function is shown in Figure 4. As Alice has observed,

[5]There are others, for example, the distribution that attributes probability one to the ball being *exactly* at the point of intersection of the parabola with the diagonal. But the

it favors the regions near the edges. Call this invariant distribution μ.

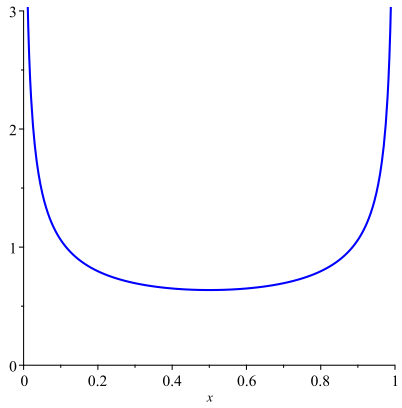


Figure 4: Density function for the invariant distribution μ.

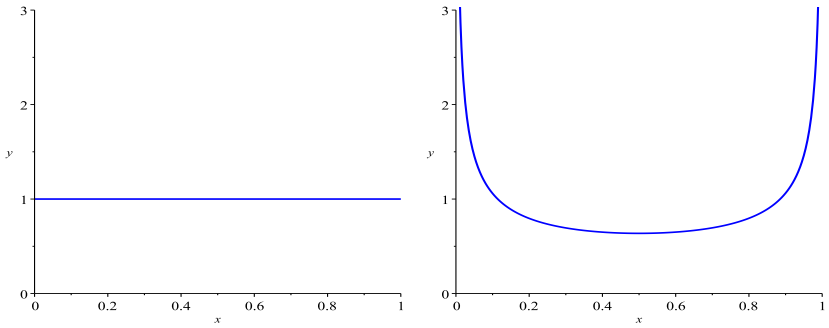
If we consider various probability distributions over initial conditions and ask what they entail about probabilities for conditions at later times, we find that, for a wide class of distributions over initial conditions, the probabilities ascribed to states of affairs only a few iterations into the future closely approximate those of the invariant distribution μ. For any "sufficiently nice" distribution over initial conditions, this approximation gets closer, without limit, as one looks farther into the future.

For example, suppose that Bob adopts a uniform distribution over initial conditions. The density function for the probabilities this bestows on states of affairs 5 iterations into the future is shown in Figure 5(a). In Figures 5(b) and (c) we see the effect of 5 iterations on other density functions for probabilities of initial conditions.

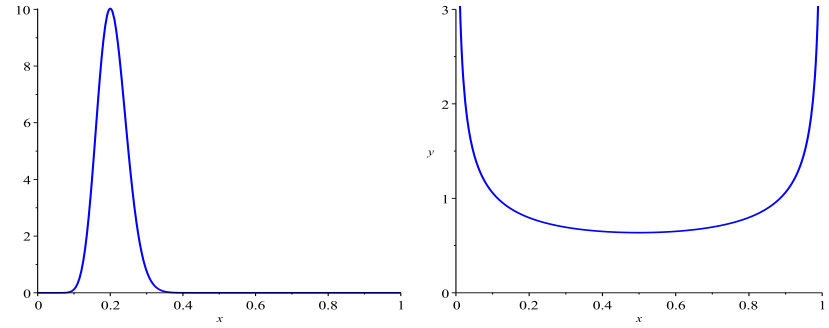
There's a theorem here: one can prove that, provided the probability distribution over initial conditions is represented by a density function that is not "too wiggly" (again, see appendix for the exact condition), then, for large n, what it says about the position of the ball n iterations into the future will be approximated by the invariant distribution μ, and, moreover, one can put bounds on how much it can depart from μ in terms of the wiggliness of the density function that yields probabilities over initial conditions; see Corollary 1.2 of Theorem 1 in the appendix.

---

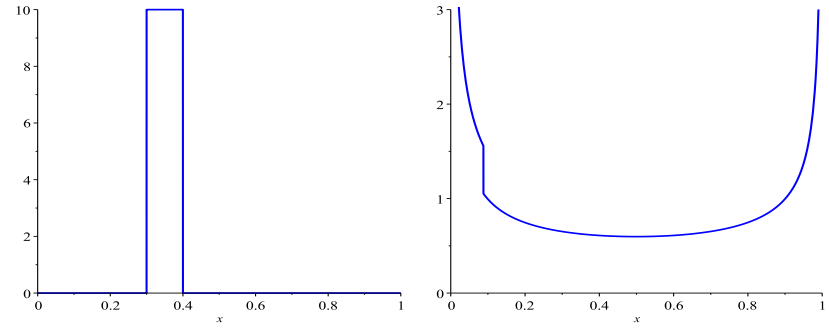
invariant distribution we're concerned with is the only one that assigns probability zero to all sets of Lebesgue measure zero, and hence can be represented by a density function.



(a) Uniform initial distribution



(b) An initial distribution peaked near 0.2



(c) Initial distribution uniform on the interval $[0.3, 0.4]$

Figure 5: The effect of five iterations of the parabola gadget on various input distributions.

Though we are not invoking a Principle of Indifference to mandate a unique credences about initial conditions, it is reasonable to expect that, even if Alice and Bob know little about the process that sets the initial conditions, they should expect there to be *some* imprecision in

the process. This will be reflected in a credence function that is represented by a continuous density function that doesn't vary excessively quickly as one moves across the diagonal. How quickly is excessively quickly depends on what a reasonable person could believe about the sorts of processes that could set initial conditions at the time the gadgets are set running.

For a gadget that has been running sufficiently long, probabilities about the position of the ball are largely independent of what one takes to be an appropriate measure over initial conditions, and one can use μ to calculate these probabilities. In this sense the dynamics of the gadget pick out μ as a "natural measure" for gadgets that have been running for a while. This is not because μ is singled out as a natural measure over initial conditions; we might be rather vague about what to think about initial conditions. It is, rather, because judgments about probabilities about the state of a gadget that has been running for a while are largely (though, of course, not entirely) independent of probabilities of initial conditions.

To adopt Bernoulli's phraseology: things like the ball being, after several iterations of the evolution, in a 10 cm. interval near the center do *not* occur with equally facility as the ball being in an interval of the same length near an edge. On a wide variety of ways of measuring sizes of sets of input states, including, but not restricted to, a measure uniform in distance along the diagonal, there is a larger measure of initial conditions that put the ball 10 cm. from the right edge than of those that put it in an interval of 10 cm. centered on the midpoint of the gadget.

4 Sensitive dependence as a source of predictability

Suppose I have $1,000$ gadgets, each of which has been running independently for some time. Consider "macrostates" of this system: we partition the width of the box into intervals of 10 cm., and specify, for each interval, how many of the balls, out of the $1,000$ gadgets, lie within that interval.

Suppose, now, I ask you to make a prediction about the macrostate. You get to choose between two propositions about the macrostate.

A'. More of the balls will be in region A than in region B.

B'. More of the balls will be in region B than in region A.

A naïve application of the Principle of Indifference of the sort favoured by Bob, on which, for each gadget, equal intervals of the diagonal are equally likely, yields a measure on which the set of states that make B' true is vastly larger than the set that makes A' true. But, as, we have seen, we should not use such a measure for gadgets that have been running for a while.

If you grant the reasoning of the previous section, then you should, for each of the gadgets, regard A as about $8/5$ times more likely than B. Moreover, the evolution of the gadgets will tend to erase correlations between initial states of distinct gadgets, so you should take probabilities regarding one gadget to be independent of probabilities regarding any other gadget. Given probabilities satisfying these conditions, it is virtually certain that the number of gadgets with balls in A will be close to $8/5$ times the number of gadgets with balls in B, and so you can be virtually certain that more of the gadgets will have their balls in A than in B. You should regard A' as overwhelmingly more probable than B'.

Let M be the probability measure over the set of $1,000$ gadgets on which each gadget is independently endowed with the invariant measure μ. On measure M, the set of states in A' is vastly larger than the set of states in B'. Moreover, any measure over initial conditions that is not too crazy will tend to agree, to a close approximation, with M about probabilities of states of affairs after a few iterations. This means that on any measure over initial conditions of the collection of gadgets that isn't too crazy, the set of initial conditions that lead, after 10 or so iterations, to states in A' is vastly larger than the set of states that lead to states in B', and, for any such measure, we can use M to compute approximately how much larger. Again: this isn't because M is favoured as a "natural measure" over initial conditions; this is a conclusion that is largely (though, of course, not entirely) independent of choice of measure over initial conditions.

In this way, we get predictability, with near certainty, of certain aspects of the state of a system consisting of a large number of parabola gadgets that have been running for an appreciable time, not *in spite of* sensitive dependence on initial conditions, but *because of it*.

It is common to distinguish between two types of regularity: regularity attributable to deterministic physical laws, and statistical regularity, regularities arising from aggregate behaviour of a large number of individually unpredictable events. Schrödinger, in *What is Life?*,

argued that all predictability is of the latter sort.

> Only in the co-operation of an enormously large number of atoms do statistical laws begin to operate and control the behaviour of there *assemblées* with an accuracy increasing as the number of atoms involved increase. It is in this way that the events acquire truly orderly features (Schrödinger, 1992, 10).

He's right about that. Whenever we make an accurate prediction, we make use of a miniscule fraction of the variables that in principle could be relevant to the outcome. Understanding how this happens requires probabilistic reasoning, even when the underlying physics is deterministic.

5 Invertibility

So far we have talked only about probabilities over initial conditions, that is, over the state of the gadget when it is set running, and their implications for probabilities for future states. Some readers will be wondering about how things might go in the other temporal direction.

Given what has been said before, readers may be forgiven for thinking that, since each position on the diagonal (except the point at which diagonal and parabola intersect) can be the result of two previous positions, the evolution of the parabola gadget is not invertible. However, there is a detail that I have so far not mentioned. In addition to the moving ball, there is a pointer that shuttles back and forth along the bottom edge of the gadget. Call the distance (in meters) of the pointer from the left edge, z. Its value changes as follows. If, at time n, the ball is to the left of center (or exactly on center), in the next iteration the value of z is halved; that is, the new position of the pointer is a distance $z/2$ from the left edge. If, at time n, the ball is to the right of center, the position of the pointer is a distance $z/2$ from the right edge. Thus, the position of the pointer at time $n+1$ carries information about the position of the ball at time n. If, at time $n+1$, the pointer is in the right half, then the ball, at time n, was also in the right half, and, if at time $n+1$, the pointer is in the left half or at the center, the the ball, at time n, was in the left half or at the center. Each position of the pointer at time $n+1$ can be reached from one and only one position at time n.

At any given stage of the evolution of the gadget, say, the nth stage, the value of z (in meters) will be a number between 0 and 1 whose binary expansion is a sequence of 1s and 0s, the first n places of which encode (starting with the most recent and going backwards) the ball's history of being to the right or left of the half mark. Thus, from precise values of the position of the ball (call it x), and of the pointer, we can reconstruct the past history of the gadget.[6] A probability distribution over the state (x, z) at some time uniquely determines probabilities for all earlier and later states, so long as the gadget runs undisturbed during the interim.

Forward evolution of the gadget leads to a situation in which the probability distribution for x is closely approximated by μ, on which the ball is equally likely to be on either side of the diagonal. Thus, after sufficiently many iterations, any information about the past of the gadget gets buried very deeply in the fine details of the distribution of z, and, for any interval $[a, b]$, the probability that z is in that interval approaches the length of the interval. The uniform distribution over z is an attractor distribution.

Let ρ be the probability distribution on which x is distributed according to μ and z is uniformly distributed, independently of x. This is an attractor distribution over the state space of the gadget. It can be proven (see appendix) that any probability distribution that can be represented by a density function converges towards this equilibrium distribution, in the sense that, for any measurable subset A of the unit square, the probability that the state (x, z) at time n will be in A approaches $\rho(A)$, as n becomes large.[7]

Evolution of the gadget does not, however, tend to smooth out probability density functions for z. Suppose, for example, that you are sure that initially the ball is to the left of center. Then, one step

[6] That is, if the ball starts out on the left side, and its history consisting of intitial state and ten iterations is (writing R for right and L for left) is $LRLLRRRLLLL$, the value of z after ten iterations has binary expansion .00001110010..., with the remaining digits determined by the initial value of z.

[7] Resist the temptation to reverse the order of the quantifiers! For any sufficiently nice input distribution, for any measurable set A and any $\varepsilon > 0$, there exists N such that, for all $n > N$, the probability that the state at time n is in A is within ε of $\rho(A)$. However, if the input distribution differs from ρ, and there is a set B such that the probability that the initial state is in B differs from $\rho(B)$ by some amount δ, then, for any n, no matter how large, there will be some set B_n such that $\rho(B_n) = \rho(B)$ and the probability that the state at time n is in B_n differs from $\rho(B_n)$ by δ.

The convergence of measures we're talking about is called *weak convergence*.

in, z will also be to the left of center. After two steps, z must be in either $[0, 1/4]$ or $[3/4, 1]$, and after 3, in $[0, 1/8]$ or $[3/8, 5/8]$ or $[7/8, 1]$. After a large number of steps, the support of the probability distribution for z will be highly fragmented, in such a way that any interval that is not too small will be half-covered by this support. This is the way that the distribution of z converges towards a uniform distribution. As proven in the appendix, it tends to go at a slower rate than the convergence of x towards its equilibrium distribution.

Since the dynamics are invertible, we can also back-evolve probability distributions. In the reverse direction we also get convergence towards the equilibrium distribution. The condition for convergence in the backwards direction is that the density function for z not be too wiggly.

Whereas forward evolution tends to turn probability density functions over x into ones that closely resemble the density function for μ and tends to complicate density functions for z, backwards evolution tends to smooth out density functions for z and complicate density functions for x. For example, consider a probability distribution that yields certainty, at some time $t+n$, that the ball is to the left, and ask what probabilities over states of affairs at time t could lead to such a thing. Because of convergence towards the equilibrium distribution in the forward direction, we know that it will have to have a very wiggly density function for x. Figure 6, shows, by way of example, the density function that results from back-evolving by 5 steps a distribution that is uniform in z and, in x, uniform over the left half of the diagonal. It is very wiggly, but, on a coarse-grained level, approximates the invariant distribution μ, in that intervals that are not too small are accorded roughly the same probability by this distributions as by μ, whose density function is also shown in Figure 6, for comparison.

To sum up:

- The measure ρ is invariant under evolution.
- It is also an attractor distribution in both forward and backwards directions. Given a probability measure over the state of the system at some time t:
  - As long as the density function for the value of x at time t is not too wiggly, the probability that the state of the system being in a set A at later time $t + n$ is approximately equal to $\rho(A)$ for large n.
  - As long as the density function for the value of z at time t is

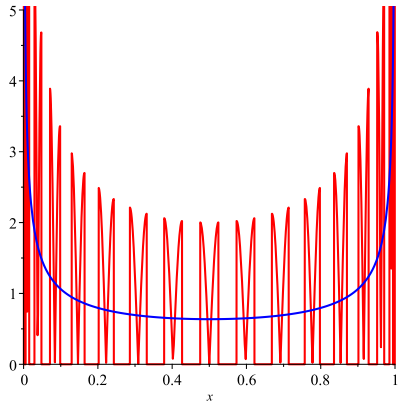


Figure 6: 5-step backwards evolute of a density function uniform over left half of the diagonal.

> not too wiggly, the probability that the state of the system being in a set A at earlier time $t-n$ is approximately equal to $\rho(A)$ for large n.

At first glance, convergence towards an equilibrium distribution in both forward and reverse directions might seem paradoxical, perhaps contradictory. Suppose, for example, we have, at time t, a distribution that is very different from the equilibrium distribution. Evolve it forward n steps, far enough into the future that the evolved distribution is a good approximation to the equilibrium distribution. If we back-evolve this n steps, will we not get something that approximates the equilibrium distribution, instead of recovering the distribution we started with?

The attentive reader will already know how to resolve this apparent paradox. The forward-evolved distribution will incorporate detailed information about the history of the system, and this will be represented by a very complicated distribution for z. The density function for z yielded by this forward-evolved distribution will be sufficiently complicated that it will take more than n steps to back-evolve it into anything like a uniform distribution.

One can say the same with the temporal orientations reversed, of course. Suppose you know that at time t, the ball is in the left side, and that the gadget had been evolving freely for a large number n steps prior to that. This would mean that you are sure that at time $t-n$, the ball was in the highly fragmented subset of states that evolve,

in n steps, into the left side of the gadget. This would be represented by a density function over x at time $t - n$ that is so complicated that forward evolution by n steps is not sufficient to bring it into an approximation of the invariant distribution μ.

To sum up: if you have detailed knowledge of the past of a time t, of the sort obtained by some other means than taking the state at t and back-evolving it, it gets encoded in a probability distribution over the state at time t via a complicated distribution for z, which renders the backwards convergence result inapplicable. Similarly, if you have detailed knowledge of the future of a time t, of the sort obtained by some other means than taking the state at t and forward-evolving it, then this would be encoded in a probability distribution over the state at time t with a complicated distribution for x, which renders the forward convergence result inapplicable.

6 Invariant distributions as surrogates

Take a probability distribution over conditions at some time t, with probabilities for x that are represented by a none-too-wiggly density function. Let n be a number of steps that is sufficiently large that probabilities for conditions n steps to the future of t are closely approximated by the equilibrium distribution. If the distribution for conditions at time t is very different from the equilibrium distribution, the distribution over z at time $t + n$ will be very complicated in its details, though it will approximate the equilibrium distribution at a coarse-grained level. These fine details will, however, be largely irrelevant for calculating probabilities over conditions at times to the future of time $t + n$, and, for the purposes of such calculations, we can replace the complicated distribution over conditions at $t + n$ by the equilibrium distribution.

To the extent that, at time $t + n$, details of the system's past have become irrelevant for its future behaviour, we can discard information about its past and use the equilibrium distribution as a surrogate for the more complicated one that encodes information about the past. It would, of course, be a mistake, when making retrodictions, to discard information about the past and to back-evolve a smoothed-out distribution.

For beings such as ourselves, who have access to records of the past but, when it comes to predicting the future, typically can do no

better than to take the current conditions and forward-evolve them, there will be an asymmetry in the invocation of convergence results. We can use a simple distribution as a surrogate for a more complicated one when it comes to predictions, insofar as the information discarded is irrelevant for future predictions, but, when it comes to retrodictions, this would be nothing sort of madness, as we would be discarding relevant information.

7 Introducing uniform distributions

The dynamics of the parabola gadget pick out a measure that is appropriate to use when making predictions concerning the future of a gadget that has been evolving freely for some time. Though uniform in z, it is not uniform in x.

Of course, whether or not a distribution is uniform depends on the variables used to characterize the state. The state of a system may equally well be represented by a different choice of variables. If one is enamoured of uniform distributions, one can indicate positions along the diagonal via a new variable, u, defined by

$$x = \sin^2(\pi u/2)\,. \tag{1}$$

On the invariant distribution μ, u is uniformly distributed: equally sized intervals of u have equal probability. As a result, when working with probability distributions over the state of the gadget, it can be more convenient to work with (u, z) rather than (x, z).

8 The "empirical way"

In Chapter IV of Part IV of *Ars Conjectandi*, Jacob Bernoulli, having shown the reader how to calculate various probabilities using combinatorics, given an equiprobable partition of events, remarks,

> It was shown in the preceding chapter how, from the numbers of cases in which arguments for things can exist or not exist, indicate or not indicate, or also indicate the contrary, the probabilities of things can be reduced to calculation and evaluated. From this it resulted that the only thing needed for correctly forming conjectures on any matter is to determine the numbers of these cases accurately and then

> to determine how much more easily some can happen than others. But here we come to a halt, because this can hardly ever be done. Indeed, it can hardly be done anywhere except in games of chance. The originators of these games took pains to make them equitable by arranging that the numbers of cases resulting in profit or loss be definite and known and that all the cases happen equally easily.[8] But this no means takes place with most other effects that depend on the operation of nature or on human will (Bernoulli 2006, 326, from Bernoulli 1713, 223).

But all is not lost. There is another way to estimate the chances of things,

> Nevertheless, another way is open to us by which we may obtain what is sought. What cannot be ascertained a priori, may at least be found out a posteriori from the results many times observed in similar situations . . . (Bernoulli 2006, 327, from Bernoulli 1713, 224).

Suppose that you did not know the exact dynamics of the parabola gadget, or, even if you did, were unable to show that the distribution ρ plays the role of an attractor. Suppose, however, that you strongly suspected that there was *some* attractor distribution towards which any probability distribution over initial conditions that could represent the credences of a reasonable agent evolved. You wouldn't know what that distribution was, but you could gain information about empirically. Let a large number of gadgets evolve for a while. Divide the diagonal into bins (not too small). Observe the positions of each of the balls, and count how many balls are in each bin. You know that the credences of any reasonable agent, evolved forward, would yield credence close to one that the relevant frequencies of balls in bins closely matched the probabilities ascribed to the bins by the attractor measure. Presumably, your credences about initial conditions are those of an agent you regard as reasonable. You should, therefore, ascribe high credence to the proposition that the observed frequencies closely approximate the attractor probabilities, and use the evidence to adjust your credences about what those attractor probabilities are. In this way, hypotheses about the attractor probabilities can be tested by experiment.[9]

[8] "... casus hi omnes pari facilitate obtingere possent."

[9] See Myrvold (2012) for a Bayesian treatment of this reasoning.

9 Status of the input distributions

I have claimed, and demonstrate in the appendix, that a wide range of probability distributions over initial conditions of the parabola gadget yield convergent probabilities for positions of the ball at later times. The dynamics map probability distributions at one time to distributions at other times, but this map requires probabilities as inputs. With no probabilities in we get no probabilities out. What, then, is the status of the input distributions we have invoked?

Following a suggestion of Savage (1973), we have been talking as if these are credences, or subjective degrees of belief. The distributions that result from evolving credences about states of affairs at time t_0, via the actual dynamics of a physical system (dynamics that might be unknown, or imperfectly known, to an agent, who might not be able to do the calculation even if the dynamics are known) are things that partake of both epistemic and physical considerations. There is an epistemic aspect, as some uncertainty about the state of the system is required. But they need not be the actual credences of any agent, because, as mentioned, an agent might not know what the result is of evolving her credences via the actual dynamics. However, in the sorts of cases we're interested in, this evolution will tend to minimize individual differences between agents' credences, and the values to which probabilities converge are determined by the dynamics of the system. These sorts of probabilities have been called *almost objective probabilities*. To emphasize that they have both epistemic and physical aspects, I have elsewhere called them *epistemic chances* (Myrvold, 2012).

All we need is some uncertainty, perhaps of a vague degree, in the agents' knowledge of initial conditions, plus some (perhaps vague) sense of a range of credence functions that are reasonable, in light of that uncertainty, and we're off and running. This need not be a purely subjective matter; judgments about what sorts of credences about initial conditions are reasonable are based on judgments about what sorts of processes there are that could lead up to those conditions.

For some systems, a classical treatment will be inadequate, and our discussion will have to be cast in terms of quantum mechanics. Such a treatment will run in much the same way. We never know for certain the precise quantum state of system. What we will want from the quantum evolution will be that it take a wide variety of initial quantum states into ones whose restriction to macrovariables

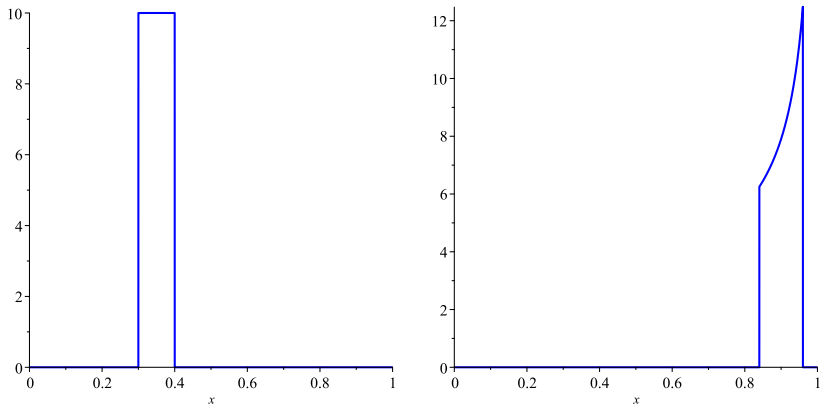
is roughly the same. (Because we're not requiring convergence of quantum states over the full set of variables, this is possible without violation of unitarity.) Thus, even if it can be argued that all probabilities in statistical mechanics, even classical statistical mechanics, have their source in quantum mechanics, convergence results of the sort we've been discussing will play a central role.

If one takes seriously (and I think that we should) the thought that the fundamental laws of physics are not deterministic, but stochastic, along the lines of a dynamical collapse theory, then the *dynamics alone* will place limitations on how much one could know about the state of a system that has been evolving for a while, because the dynamics alone will produce a range of possible states from one and the same initial state. Nonetheless, if the stochastic dynamics is going to produce convergence towards certain probability distributions over quantum states, results of that sort are likely to take the form of examination of behaviour of a range of quantum states under deterministic evolution, with the role of the collapse dynamics being that of providing the requisite uncertainty in the state of the system at a given time.

10 On the way to equilibrium: partial equilibration and autonomous equations

Consider a probability distribution that is initially concentrated on some subinterval of the diagonal, such as the one whose density function is depicted in Figure 7a; this one is confined to the interval $[0.3, 0.4]$, and is uniform, in position along the diagonal, on that interval. It takes several iterations of the gadget to spread this distribution over the full width of the diagonal. But something interesting happens in the meantime. Take a look at Figure 7b. There we see the one-step evolute of the distribution of 7a. It is confined to the interval $[0.84, 0.96]$, but, on that interval, it closely approximates the restriction of μ to that interval.

Call this phenomenon *local equilibration*. The probability distribution is nothing like the equilibrium distribution, as the position is confined to a sub-interval. But, subject to that constraint, it comes near to being as much like the invariant distribution as it could be, while satisfying that constraint. We can, without significant loss, replace the probability distribution for x by one that is the restriction of the invariant distribution to that interval.



(a) (b)

Figure 7: Local equilibration

The fact that probability distributions for x approach the invariant distribution rapidly has implications for the evolution of the other variable z, which, as already mentioned, equilibrates more slowly. Suppose we were concerned about tracking the evolution of z, and were either uninterested in or had no access to the value of x. This would pose a problem. Since the value of z on the $(n+1)$th iteration depends, not only on z, but also on x, the probability distribution for z at the $(n+1)$th stage depends on the probability distribution for the full state (x, z) at the nth stage.

For *certain kinds* of distributions, however, we can get an autonomous equation that gives probabilities for future values of z in terms of its present value. If the probability distribution for x is symmetric around the mid-point, then, with probability $1/2$, z_{n+1} is $z_n/2$, and, with probability $1/2$, it is $1 - z_n/2$. This, in turn, permits us to get an equation that produces a probability distribution for z_{n+1} from a distribution for z_n.

Not all probability distributions for x have this feature. However, as we have seen, sufficiently nice probability distributions for x rapidly approach the invariant distribution, which has the requisite symmetry. For such initial distributions, after a few iterations, the autonomous equation that yields probabilities for z_{n+1} in terms of the value of z_n will be a good approximation to what one would obtain using the probability distribution over the full state space.

What makes this work is separation of time scales. The variable x

changes its value rapidly—nearby values of x, in one iteration, double their separation, whereas, for almost any fixed value of x, nearby values of z approach each other; values of x are *unstable*, compared to values of z. As a result, evolution tends to smooth out density functions for x and complicate density functions for z. The fact that x is unstable compared to z is reflected in rapid approach towards the invariant probability distribution for x, compared to the rate of approach of z towards its invariant distribution.

This has an analog in real physical systems. Consider, for example, a helium balloon made of rubber. As every child knows, the balloon is permeable to helium and to air, and after a while, the balloon will reach an equilibrium state in which the gas inside it is the same pressure and composition as the atmosphere outside. But this happens slowly, and, in the meantime, we can treat the system as one in which there is helium gas at high pressure inside the balloon and ordinary atmospheric mix of gases outside. If, on a warm sunny day, you take a balloon that has been outside into an air conditioned building, the contents of the balloon will rapidly come to the same temperature as the air in the room, and we will be able, over time scales short on the time scale of leakage of helium from the balloon, to treat the gas inside the balloon and outside as each being in a state of thermal equilibrium, at the same temperature but unequal pressures. This phenomenon is not uncommon; all is needed is a separation of time scales, in that some variables equilibrate fairly quickly, others, more slowly.

A striking example can be found in the treatment of Brownian motion. As mentioned in the introduction, the standard treatment invokes the Langevin equation, a stochastic equation for the velocity of the Brownian particle. The force terms in this equation are the ones that would result from treating the molecules of the fluid as being in thermal equilibrium, even if the state of the system as a whole, which includes the velocity of the Brownian particle, has not had time to equilibrate. The rationale for this is that equilibration of the fluid molecules is faster than that of the Brownian particle, which acts, as it relaxes towards equilibrium, as if it is immersed in a fluid in equilibrium.

It is a feature of this treatment of Brownian motion that probabilities concerning the future velocity of the particle depend only on the present velocity, whether the process that gave rise to that velocity was slowing, through friction, of a higher speed, or via a fluctuation

from a lower speed. The rationale is that, though the detailed state of the fluid at any given time will contain traces of the Brownian particle's past history, any molecule that interacts with the particle will soon interact with other molecules and these traces will become distributed through the fluid in a complicated way, and will become irrelevant to the future behaviour of the particle, permitting us to treat it, for purposes of predicting future movements of the particle, as if it were newly introduced with its current velocity into a fluid in equilibrium. This treatment yields predictions that are in very good agreement with experiment, and this, in turn, gives us evidence that we were correct to think that, whatever traces there may be in the surrounding medium of the Brownian particle's previous state of motion, they are largely irrelevant to its subsequent behaviour.

It is considerations such as these that underwrite the *method of projections*, discussed in §6 of Wallace (2015). This is used to obtain an equation governing a probability distribution over a reduced set of variables, typically a set that falls far short of the full set of variables. Since, in general, future values of each variable depends on the present values of *all* of them, it is not, as Wallace says, immediately clear how this can be done. The key is local equilibration. If probability distributions over initial conditions evolve rapidly (compared to the time scales of interest) into ones that are such that, for the purpose of studying the evolution of the limited set of variables of interest, the remaining variables can be treated as if they are distributed according to some equilibrium distribution, then we can treat the system as if, at any given time, those remaining variables *do* have the equilibrium distribution, and evolve the probability distribution for the variables of interest accordingly.

We have seen this explicitly in the case of the parabola gadget. A full probability distribution over the state space will involve complicated relations between the precise value of x and the precise value of z. However, for the purposes of forward-evolving z, all that is needed is the current value of z and very limited information about the distribution of x; all that matters is the very coarse-grained probability that tells us probabilities for x being in the left or right half. Moreover, initial distributions for x rapidly evolve into ones for which these probabilities are equal, and so we can get an equation that relates future probabilities of z to present ones.

11 Prediction and retrodiction

Suppose that, for some physical system with dynamics that are invariant under time-reversal, you are convinced that something like the separation of times-scales discussed in the preceding section is in operation, and that, for the purposes of prediction of macrovariables, any reasonable probability distribution can be replaced by one that is as close to uniform as possible, given the current macrostate. This conviction might be based on calculation, on empirical evidence, or on a combination of the two.

Suppose, now, that you run across a system of that type which you have good reason to believe has been evolving in isolation for some time, and find it in some nonequlibrium macrostate (you may think, if you like, of the familiar example of a thermos bottle containing warm water and some ice cubes). What should you predict, and what should you retrodict, about the system?

For prediction, we've already said that you are convinced that taking a distribution that is as close to uniform as possible, given the current microstate, and evolving it forward, is an effective strategy. One might be tempted to conclude that reversible dynamics, together with the effectiveness of the strategy in the forward direction, either suggests or even *requires* you to take it to be an effective strategy in the reverse direction, in which case you should back-evolve the smoothed-out distribution, and conclude that the system was probably closer to equilibrium in the recent past.

This is too quick, and, in fact, is justified only if you take the role of the uniform, invariant distribution to be one that guides your expectations about what is typical and atypical, instead of the role for which you have a rationale, namely, as a *surrogate* for a distribution that represents your belief state about the system. To see this, let us undertake a Bayesian calculation. Let E be the evidence that the system is, at time t_0, in the observed nonequilibrium state. Let H_{non} be the hypothesis that the system was in some nonequilibrium state that is such that application of the recipe of forward-evolving a smoothed distribution over that macrostate accords high probability to E: *e.g.* a thermos with larger ice cubes and warmer water than observed. Let H_{eq} be the result of back-evolving the observed macrostate; this will be a state closer to equilibrium than the observed macrostate, for instance, a state with smaller ice cubes and cooler water than the

observed state. Bayes' theorem tells us that

$$\frac{Pr(H_{non}|E)}{Pr(H_{eq}|E)} = \frac{Pr(E|H_{non})}{Pr(E|H_{eq})} \times \frac{Pr(H_{non})}{Pr(H_{eq})}. \tag{2}$$

Since H_{eq} and H_{non} are hypotheses about the state at an earlier time that the observed state involved in E, we can apply the recipe of forward-evolving a smoothed probability distribution to compute $Pr(E|H_{non})$ and $Pr(E|H_{eq})$. By construction the former is large; the latter will be tiny, as it is the probability of spontaneous heat transfer from ice to warm water, resulting in spontaneous freezing. Thus, the first factor on the right-hand side of (2) is huge. Therefore, $Pr(H_{non}|E)$ will be much, much larger than $Pr(H_{eq}|E)$, and you should regard it as much more likely that the system reached its observed state from a further-from-equilibrium state, *unless* you take the second factor to be miniscule, that is, unless you take the prior probability of H_{eq} to be enormously larger than the prior probability of H_{non}.

That is, if you are convinced, by experience or otherwise, that evolving a smoothed distribution over the system's macrostate works well in the forward direction, you should apply it also in the reverse direction only if you are already convinced that your priors about the recent past should be set according to a measure according to which equilibrium states are typical and nonequilibrium states mind-boggling improbable.

12 Comparison with real systems

The parabola gadget is a toy model, but it shares relevant features with real systems.

For an isolated classical system, as is well-known, there is at least one invariant measure, namely the microcanonical measure. For large, complicated systems, we expected sensitive dependence on initial conditions; this means that evolution will tend to separate points along along certain dimensions and, because measure is preserved, there must be a corresponding compression along other dimensions. That means that density functions for some variables will tend to smooth out and others, to become more complicated, just as we have seen in the parabola gadget.

There is a disanalogy between the parabola gadget and real systems, in that, unlike most (all?) real systems to which statistical

mechanics is applied, the dynamics of the gadget are not invariant under time reversal.[10] This disanalogy is less important than might seem, as the gadget's lack of T-invariance plays no role in the equilibration result, which, as already mentioned, works in both temporal directions. Moreover, its dynamics *are* invariant under a combination of time reversal and a transformation that swaps the coordinates u and z. This is reminiscent of the fact that, though, strictly speaking, fundamental laws are not invariant under time reversal, they are, in the standard model, invariant under the combination of time reversal, charge conjugation, and parity inversion (CPT).

The dynamics of the parabola gadget are ergodic and strong mixing. This is not required for the sorts of conclusions we are interested in, as, for real systems, we will not be interested in convergence in probability for all degrees of freedom of the system, but only a limited subset of macroscopically measurable variables. Nor is it sufficient, for our purposes, that a system be mixing or have some other place in the ergodic hierarchy, as results of this sort concern only long-run limiting behaviour, and we are interested in knowing what happens at finite times.

Unlike the parabola gadget, real macroscopic systems systems have a large number of degrees of freedom. This makes them hard to deal with analytically, but allows one to invoke considerations of the behaviour of limited subsystems of systems with a large number degrees of freedom. There are a number of theorems, of some generality, giving convergence results for systems of this sort. These apply, both to a large isolated system which is such that we have access only to a small number of macroscopic variables, and (what amounts to the pretty much the same thing) to a small system interacting with a large and complicated environment (take, in such a case, the system's degrees of freedom as a small subset of the degrees of freedom of the system + environment).

Recent work along these lines has, understandably, focussed on equilibration of quantum systems, because, ultimately, systems we treat classically are to be treated in terms of quantum mechanics. Linden et al. (2009) have shown that, for a very broad class of Hamil-

[10]The reason for the question-mark is that, when weak-force interactions are taken into account, we must acknowledge that the fundamental laws are not T-invariant. This, however, is irrelevant to questions regarding thermodynamic asymmetry, as the fundamental laws, as we currently have them, are CPT-invariant, and thermodynamic asymmetries are also CPT-asymmetries.

tonians (namely, those with nondegenerate energy gaps), the reduced state of a small subsystem of a large quantum system will equilibrate, provided only that the state of the large system contain a large number of energy eigenstates. Also of interest is the result due to Goldstein et al. (2010), which demonstrates approach to macroscopic equilibrium for arbitrary initial states and "typical" Hamiltonians. For recent work along these lines, see Goldstein et al. (2015) and references therein.

13 Conclusion

Taking standard measures as if they are required by a Principle of Indifference raises several problems. One is the question of justification: why this measure, rather than some other? If this question could be satisfactorily answered, this would raise a more severe problem in that, applied out of equilibrium, the standard measures are spectacularly wrong, and they give rise to disastrous retrodictions.

Dynamical considerations come to our aid in picking out an appropriate measure. Moreover, when the rationale for the use of this measure is understood, we see that its legitimate employment is much more restricted than a Principle of Indifference would suggest. In particular, we are not obliged to regard the early state of the universe and virtually everything we see as extraordinarily improbable. As a corollary of this, there is no incentive whatsoever to take the appropriate measure over the current state of affairs to be one that is invariant under velocity reversal, and hence do not end up saddled with the disastrous retrodictions a measure of that sort would engender.

We can apply considerations of this sort, not only to the end point of the process of equilibration, but also to intermediate steps. What is required is a difference in time scales of relaxation of input distributions towards equilibrium distributions. If some variables relax more quickly than others, we can obtain autonomous equations for probability distributions of the more slowly changing variables.

Of course, this approach leaves work to do, work that is an active area of research in nonequilibrium statistical mechanics: that of showing that, for something like real systems, the requisite convergence results hold. Rather than simply taking for granted the positive outcome of this endeavour, philosophers of statistical mechanics would do well to pay more attention to research along these lines.

14 Appendix

In this appendix, we make and prove precise the convergence results alluded to above. For related results, see Engel (1992) (in particular, Theorem 3.9a and §3.3.6).

Let (x_n, z_n) be the state of the gadget at stage n of its evolution. The dynamics specified in the main text gives

$$x_{n+1} = 1 - 4(x_n - 1/2)^2 = 4x_n(1 - x_n). \tag{3}$$

$$z_{n+1} = \begin{cases} z_n/2, & \text{if } x_n \leq 1/2; \\ 1 - z_n/2, & \text{if } x_n > 1/2. \end{cases} \tag{4}$$

Readers familiar with the literature on chaos theory will already have recognized (3) as the logistic map.

A probability distribution for x_0 determines distributions for each x_n, $n > 0$. Suppose that we have a probability distribution for x_0 that has density f with respect to the invariant measure μ. That is, for any measurable subset A of the unit interval,

$$Pr(x_0 \in A) = \int_A f(x)\, d\mu(x). \tag{5}$$

If g is the corresponding density with respect to Lebesgue measure on the unit interval, the two are related by

$$f(x) = \pi\sqrt{x(1-x)}\, g(x). \tag{6}$$

Thus, the invariant distribution, which has flat density with respect to itself, has density, with respect to Lebesgue measure,

$$g_\mu(x) = \frac{1}{\pi\sqrt{x(1-x)}}, \tag{7}$$

which is the function we have seen graphed in Figure 4, above.

However, it is the density f, the density with respect to the invariant measure μ, that will be of interest to us, as it is in terms of this density that we obtains bounds relevant to rates of convergence. This is a good thing; as is clear from (6), the function f varies less than the function g, and it is in terms of the variation of f that we will find our bounds. In particular, we will find bounds for distributions such that f has finite total variation (see below), but g need not have finite total variation for such bounds to apply.

It will be useful to make a change of variables. Define the variable u by

$$x = \sin^2\left(\frac{\pi u}{2}\right). \tag{8}$$

This variable is useful because the invariant distribution μ is uniform in u. That is, for any interval $[a, b]$ within the unit interval,

$$\mu(\{u : u \in [a, b]\}) = b - a. \tag{9}$$

The evolution (3) induces a corresponding map for u:

$$\sin^2\left(\frac{\pi\, u_{n+1}}{2}\right) = 4\sin^2\left(\frac{\pi\, u_n}{2}\right)\cos^2\left(\frac{\pi\, u_n}{2}\right) = \sin^2\left(\pi\, u_n\right). \tag{10}$$

This gives,

$$u_{n+1} = 1 - 2|u_n - 1/2| = \begin{cases} 2u_n, & u_n \le 1/2; \\ 2(1 - u_n), & u_n > 1/2. \end{cases} \tag{11}$$

This is the tent map. It's easy to work with because it's piecewise linear.

The invariant measure ρ is uniform in u and z. It can be shown that the evolution is *strong mixing* with respect to ρ: that is, for any measurable sets A, B,

$$\rho(A_n \cap B) \to \rho(A)\rho(B) \text{ as } n \to \infty, \tag{12}$$

where A_n is the result of applying n iterations of the evolution to A. Hopf's proof (1934, §8) that the baker's map is strong mixing applies equally well to the evolution of the gadget. From this it follows that any distribution over initial conditions that has a density with respect to the invariant measure weakly converges towards this measure. That proof doesn't, without further ado, provide information about bounds on rates of convergence, which we now investigate.

Iteration of the tent map n times produces 2^{n-1} copies of the tent, each supported on an interval of length $2^{-(n-1)}$. Each of these intervals consists of two subintervals of length 2^{-n} that are mapped linearly onto the unit interval. Let Δ_i, for $i = 1, \ldots, 2^n$, be the subinterval $[(i-1)/2^n, i/2^n)$.

Now, let us consider some measurable subset B of the unit interval, with measure $\mu(B)$, and consider its inverse image under n-fold iteration of the tent map; that is, consider the set A that gets mapped into B. Each of the subintervals Δ_i contains a subset $A_i = A \cap \Delta_i$

of measure $\mu(B)/2^n$. The probability that u_n is in B is equal to the probability that u_0 is in A, which is the sum of the probabilities of $u_0 \in A_i$ over all the subintervals A_i.

Expressed in terms of our original variable x: n-fold iteration of the logistic map partitions the unit interval into 2^{n-1} intervals of equal μ-measure $2^{-(n-1)}$, each of which contains two subintervals of measure 2^{-n} that get mapped onto the unit interval. For any measurable subset B, with inverse image A, each of these subintervals contains a subset of A of measure $\mu(B)/2^n$.

We want to investigate the probability that a distribution with a given density function f ascribes to A. We will make use of the following theorem.

Theorem 1. *Let x be a random variable that has density f with respect to some measure μ. Let A be a measurable set with the property that the range of x can be partitioned into subsets Δ_i such that, for each i,*

$$\mu(\Delta_i \cap A) = \mu(A)\,\mu(\Delta_i).$$

Let f_i^+, f_i^-, be the essential supremum and essential infimum, respectively, of f on Δ_i. Then

$$|Pr(x \in A) - \mu(A)| \;\leq\; \mu(A)\,\mu(\bar{A}) \sum_i \mu(\Delta_i)(f_i^+ - f_i^-),$$

where $\bar{A}$ is the complement of A.

From this follows a corollary in terms of the *total variation* of the density function f. Alpine hikers will find the concept of total variation intuitive. Imagine walking along the graph of the function from left to right. The total variation is the total vertical ascent plus the total vertical descent you have to do. The official definition (which is found in many textbooks of real analysis; see, *e.g.*, Kolmogorov and Fomin 1975, §9.32) is as follows.

Definition. Consider a function $g : [a, b] \to \mathbb{R}$, defined on some interval $[a, b]$ of the real line. Take any finite increasing set of numbers $a = x_0 < x_1 < ... < x_n = b$, and consider the quantity

$$\sum_{k=1}^{n} |g(x_k) - g(x_{k-1})|.$$

The *total variation* of g, $V(g)$, is defined to be the essential supremum of this quantity, over all choices of $x_0, ..., x_n$.

Obviously, a constant function has total variation zero. If a function has finite total variation, it is said to be of *bounded variation*. The density function of the invariant measure μ with respect to Lebesgue measure, given in equation (7), has unbounded total variation, but its density function with respect to itself is flat and has total variation zero. A function of bounded variation on $[a, b]$ has a finite derivative at almost all points in $[a, b]$ (Kolmogorov and Fomin, 1975, 331). If g is a continuous function of bounded variation, then

$$V(g) = \int_a^b |g'(x)|\, dx.$$

If g is piecewise continuous, we add to this the sum of the absolute values of the jumps that g makes at each of its points of discontinuity.

From the above theorem the following corollary is immediate.

Corollary 1.1. *Under the conditions of Theorem 1, if f has finite total variation $V(f)$, and if there exists δ such that $\mu(\Delta_i) \leq \delta$ for all i, then*

$$|Pr(x \in A) - \mu(A)| \;\leq\; \mu(A)\,\mu(\bar{A})\;\delta\;V(f).$$

Applied to n-fold iteration of the logistic map: each of the subsets Δ_i has measure 2^{-n}. This yields the desired convergence result for the parabola gadget, regarding distributions of the variable x.

Corollary 1.2. *Let $\{x_k\}$ be a sequence of random variables related by the logistic map (3). Let x_0 have a distribution that has density f with respect to the invariant measure μ, with finite total variation $V(f)$. For any n, partition the unit interval into 2^n intervals of equal measure $\mu(\Delta_i) = 2^{-n}$, and take f_i^+ and f_i^- to be the essential supremum and infimum, respectively, of f on Δ_i. Let A be a measurable set, with complement $\bar{A}$. Then*

$$\begin{aligned} |Pr(x_n \in A) - \mu(A)| \;&\leq\; \frac{\mu(A)\,\mu(\bar{A})}{2^n} \sum_{i=1}^{2^n} \left(f_i^+ - f_i^-\right) \\ &\leq\; \frac{\mu(A)\,\mu(\bar{A})}{2^n}\, V(f). \end{aligned}$$

We now prove Theorem 1.

Proof. let χ_A be the characteristic function of A,

$$\chi_A(x) = \begin{cases} 1, & x \in A; \\ 0, & x \notin A. \end{cases} \tag{13}$$

Let $\lambda = \mu(A)$.

$$\begin{aligned} Pr(x \in A) - \lambda &= \int f(x)\,(\chi_A(x) - \lambda)\,d\mu(x) \\ &= \sum_i \int_{\Delta_i} f(x)\,(\chi_A(x) - \lambda)\,d\mu(x). \end{aligned} \tag{14}$$

Since $\mu(\Delta_i \cap A) = \lambda\,\mu(\Delta_i)$ for all i,

$$\int_{\Delta_i} (\chi_A(x) - \lambda)\,d\mu(x) = 0, \tag{15}$$

and so, for any numbers $\{a_i\}$,

$$Pr(x \in A) - \lambda = \sum_i \int_{\Delta_i} (f(x) - a_i)\,(\chi_A(x) - \lambda)\,d\mu(x). \tag{16}$$

Therefore,

$$|Pr(x \in A) - \lambda| \leq \sum_i \int_{\Delta_i} |(f(x) - a_i)\,(\chi_A(x) - \lambda)|\,d\mu(x). \tag{17}$$

Take

$$a_i = \frac{1}{2}\left(f_i^+ + f_i^-\right). \tag{18}$$

Then, for almost all $x \in \Delta_i$,

$$|f(x) - a_i| \leq \frac{1}{2}\left(f_i^+ - f_i^-\right), \tag{19}$$

and (17) yields,

$$|Pr(x \in A) - \lambda| \;\leq\; \frac{1}{2}\sum_i \left(f_i^+ - f_i^-\right) \int_{\Delta_i} |\chi_A(x) - \lambda|\,d\mu(x). \tag{20}$$

Within Δ_i, the function $|\chi_A(x) - \lambda|$ is equal to $1-\lambda$ on a set of measure $\lambda\,\mu(\Delta_i)$ and to λ on a set of measure $(1-\lambda)\mu(\Delta_i)$. Therefore,

$$\int_{\Delta_i} |\chi_A(x) - \lambda|\,d\mu(x) = 2\lambda(1-\lambda)\mu(\Delta_i). \tag{21}$$

We thereby obtain the result,

$$|Pr(x \in A) - \lambda| \leq \lambda(1-\lambda)\sum_i \mu(\Delta_i)(f_i^+ - f_i^-), \tag{22}$$

which is what was to be proved. □

We have demonstrated convergence of probability distributions for x. This means, for any measurable subset A of the unit interval, the measure of the subset of the state space that consists of all (x, z) with $x \in A$ converges towards $\mu(A)$. We will now show that we have convergence of measure, not only for such sets, but for arbitrary measurable subsets of the state space.

It suffices to show that we have convergence for rectangles $[a, b] \times [c, d]$. Now, if we partition the z-axis into 2^k bins of length 2^{-k}, then, after k iterations of the gadget's evolution, which bin z is in depends only on the initial value of x and is independent of the initial value of z. Therefore, for a rectangle of the form $[a, b] \times [p/2^k, q/2^k]$, where p and q are integers, the probability that, at any time after k iterations, the state is in that rectangle, depends only on the initial value of x. Its inverse image is the set of all points such that x is in a set D, where D has measure $\mu(D) = (b - a)(q - p)/2^k$.

We can say the same of any set that is the set of all (x, z) for $x \in A$ and $z \in L$, where A is any measurable subset of the unit interval and L is a union of intervals, of total length $\|L\|$, with endpoints that are integral multiples of 2^{-k}. The inverse image of this set under k iterations is the set of all (x, z) with $x \in D$, where D is a set of measure $\mu(D) = \rho(A \times L) = \mu(A)\|L\|$. Thus, for any n, k, $(x_{n+k}, z_{n+k}) \in A \times L$ if and only if $x_n \in D$, and

$$Pr((x_{n+k}, z_{n+k}) \in A \times L) = Pr(x_n \in D). \tag{23}$$

Let $\lambda = \mu(D) = \rho(A \times L)$. Then

$$|Pr((x_{n+k}, z_{n+k}) \in A \times L) - \lambda| = |Pr(x_n \in D) - \lambda| \ \leq \ \frac{\lambda(1 - \lambda)}{2^n} \, V(f), \tag{24}$$

where f, once again, is the density function, with respect to μ, for x_0. This gives us the following convergence result for sets of this form.

Theorem 2. *Let A be any measurable subset of the unit interval, and let L be a subset of the unit interval that is a union of intervals with endpoints that are integral multiples of 2^{-k}. Let $\lambda = \rho(A \times L)$. Then, for $n \geq k$,*

$$|Pr((x_n, z_n) \in A \times L) - \lambda| \ \leq 2^k \ \frac{\lambda(1 - \lambda)}{2^n} \, V(f),$$

where f is the density function for x_0 with respect to μ.

We get convergence of the distribution of z by taking A to be the unit interval.

Corollary 2.1. *Let L be a subset of the unit interval that is a union of intervals with endpoints that are integral multiples of 2^{-k}, of total length λ. Then, for $n \geq k$,*

$$|Pr(z_n \in L) - \lambda| \leq 2^k \frac{\lambda(1-\lambda)}{2^n} V(f),$$

where f is the density function for x_0 with respect to μ.

If we coarse-grain both u and z into 2^k bins of equal width 2^{-k}, then, for any initial distribution such that f has bounded total variation, the probability of u_n being in a given bin approaches 2^{-k} as n increases, as does the probability of z_n being in a given bin, but we have tighter bounds on the probability of u_n, by a factor 2^k.

It is easy to extend this convergence result to arbitrary intervals $L = [a, b]$, to get bounds on the rate of convergence of the distribution of z. For any interval, and any k, we can find intervals $L' = [a', b']$, $L'' = [a'', b'']$, whose endpoints are integral multiples of 2^{-k}, such that $a'' \leq a \leq a'$, $b' \leq b \leq b''$, $a' - a'' \leq 2^{-k}$ and $b'' - b' \leq 2^{-k}$. Since $L' \subseteq L \subseteq L''$, we must have

$$Pr(z_n \in L') \leq Pr(z_n \in L) \leq Pr(z_n \in L''). \tag{25}$$

Let λ, λ', and λ'' be the lengths of L, L', and L'', respectively. λ' and λ'' are both within 2^{-k} of λ. We have, from (25),

$$\begin{aligned} Pr(z_n \in L') - \lambda' - (\lambda - \lambda') \leq Pr(z_n \in L) - \lambda \\ \leq Pr(z_n \in L') - \lambda'' + (\lambda'' - \lambda). \end{aligned} \tag{26}$$

and hence

$$\begin{aligned} -\left(|Pr(z_n \in L') - \lambda'| + 2^{-k}\right) \leq Pr(z_n \in L) - \lambda \\ \leq |Pr(z_n \in L'') - \lambda''| + 2^{-k}. \end{aligned} \tag{27}$$

We can apply Corollary (2.1) to L' and L'' ; by taking k and n sufficiently large, we can get bounds as tight as we want on $|Pr(z_n \in L) - \lambda|$.

15 Acknowledgments

I am privileged to have benefited from discussions these matters with a number of people over the years. I acknowledge, in particular, David Albert, Harvey Brown, Jeremy Butterfield, Alison Fernandes, Roman Frigg, Shelly Goldstein, Carl Hoefer, Jenann Ismael, Josh Luczak, Owen Maroney, John Norton, Jos Uffink, David Wallace, and Charlotte Werndl.

References

Albert, D. (2000). *Time and Chance*. Cambridge: Harvard University Press.

Bernoulli, J. (1713). *Ars Conjectandi, Opus Posthumum*. Basel: Impensis Thurnisiorum, Fratrum.

Bernoulli, J. (2006). *The Art of Conjecturing*. Baltimore: Johns Hopkins University Press. Translated by Edith Dudley Sylla.

Butterfield, J. (1996). Whither the minds? *The British Journal for the Philosophy of Science 47*, 200–221.

Engel, E. M. (1992). *A Road to Randomess in Physical Systems*. Berlin: Springer-Verlag.

Goldstein, S. (2001). Boltzmann's approach to statistical mechanics. In J. Bricmont, D. Dürr, M. Galavotti, G. Ghirardi, F. Petruccione, and N. Zanghì (Eds.), *Chance in Physics*, pp. 39–54. Berlin: Springer.

Goldstein, S., T. Hara, and H. Tasaki (2015). Extremely quick thermalization in a macroscopic quantum system for a typical nonequilibrium subspace. *New Journal of Physics 17*, 045002.

Goldstein, S., J. L. Lebowitz, C. Mastrodonati, R. Tumulka, and N. Zanghì (2010). Approach to thermal equilibrium of macroscopic quantum systems. *Physical Review E 81*, 011109.

Hacking, I. (1971). Equipossibility theories of probability. *The British Journal for the Philosophy of Science 22*, 339–355.

Hopf, E. (1934). On causality, statistics, and probability. *Journal of Mathematics and Physics 13*, 51–102.

Kolmogorov, A. N. and S. Fomin (1975). *Introductory Real Analysis*. New York: Dover Publications. Translated and edited by R. A. Silverman.

Laplace, P.-S. (1814). *Essai Philosophique sur les Probabilités*. Paris: Courcier. English translation in Laplace (1902).

Laplace, P.-S. (1902). *A Philosophical Essay on Probabilities*. New York: John Wiley & Sons. Translation of Laplace (1814).

Linden, N., S. Popescu, A. J. Short, and A. Winter (2009). Quantum mechanical evolution towards thermal equilibrium. *Physical Review E 79*, 061103.

Mazenko, G. F. (2006). *Nonequilibrium Statistical Mechanics*. Weinheim: Wiley-VCH.

Myrvold, W. C. (2012). Determistic laws and epistemic chances. In Y. Ben-Menahem and M. Hemmo (Eds.), *Probability in Physics*, pp. 73–85. Springer.

Price, H. (2002). Boltzmann's time bomb. *The British Journal for the Philosophy of Science 53*, 83–119.

Savage, L. J. (1973). Probability in science: A personalistic account. In P. Suppes (Ed.), *Logic Methodology, and Philosophy of Science IV*, pp. 417–428. Amsterdam: North-Holland.

Schrödinger, E. (1992). *What is Life? with Mina and Matter and Autobiographical Sketches*. Cambridge: Cambridge University Press.

Wallace, D. (2015). The quantitative content of statistical mechanics. *Studies in History and Philosophy of Modern Physics 52*, 285–293.

Wallace, D. (forthcoming). The logic of the past hypothesis. In B. Loewer, E. Winsberg, and B. Weslake (Eds.), *Time's Arrows and the Probability Structure of the World*. Harvard: Harvard University Press. Available at `http://philsci-archive.pitt.edu/8894/`.

Zwanzig, R. (2001). *Nonequlibrium Statistical Mechanics*. Oxford: Oxford University Press.